

Evaluating Inter-Expert Variability and Model Performance in Automated Brain Metastasis Segmentation

1/2

Anne-Charlotte Roux¹, Laurie Marchi¹, Alexis Perrot¹, Stéphanie Rudzinka¹, François Lucia¹, Brieg Dissaux¹, Thomas Theodoridis², Alexis Bombezin-Domino³, Anne Walrafen³, Sami Romdhani², Gizem Temiz³, Nikos Paragios⁴, Vincent Bourbonne^{1,5}

¹Radiology Department, University Hospital Brest, Brest, France. ²AI Engineering, Therapanacea, Paris, France. ³Clinical Affairs, Therapanacea, Paris, France. ⁴CEO, Therapanacea, Paris, France. ⁵GETBO, University of Brest, Brest, France

Purpose/Objective:

Brain metastasis segmentation is essential for accurate treatment planning in radiotherapy. However, manual segmentation remains a challenging and time-consuming task due to the variability in tumor shapes, sizes, and locations, as well as inter-expert variability. Automated segmentation models offer a potential solution by providing consistent and efficient tumor delineations. This study assessed the performance of a model designed for brain metastasis segmentation by comparing it against a reference mask generated through majority voting of annotations from five experts. In addition, we evaluated inter-expert variability through Dice similarity to contextualize the model's performance against human agreement levels.

Material/Methods:

We developed an automatic brain metastasis segmentation model utilizing a large mixed model [1] trained on multicentric data from 397 patients, sourced from both hospital practices and the BraTS challenge [2]. In a subset of 39 cases from the brain metastasis unseen test set, five experts provided annotations for individual tumors. To assess agreement on each tumor, these annotations were combined into a single mask per patient by summing their values. A reference mask was then generated by applying a majority-vote threshold to the summed mask. Tumors in this reference mask were identified through 3D connected components analysis, enabling assessment of agreement between the reference and predicted masks in terms of precision, sensitivity, and false positive rate. Additionally, an inter-expert analysis (IEV) was conducted using patient-wise Dice similarity. Dice similarity scores were calculated between all pairs of experts' contours and between all pairs of model-versus-experts. These 2 distributions were then compared using medians. In the end, all median values obtained per case were averaged over the whole cohort to obtain the end results.

Results:

The model achieved a precision of 0.87, a recall of 0.78, and a false positive rate of 0.39 when compared to the majority-voted reference mask. In the inter-expert validation IEV analysis for these cases, the average Dice similarity among experts was 0.85, whereas the Dice similarity for the predicted contours was 0.78.

Evaluating Inter-Expert Variability and Model Performance in Automated Brain Metastasis Segmentation

Conclusion:

Our study suggests that the automatic brain metastasis segmentation model performs comparably to expert annotations, demonstrating potential as a reliable tool in clinical workflows. However, to more accurately evaluate both IEV and model performance relative to human annotations, future analyses could benefit from the creation of a consensus mask.

Keywords: Brain metastasis, segmentation, AI